Automatic Analysis of Affect and Membership in Group Settings

Wenxuan Mou[1], Hatice Gunes[2], Ioannis Patras[1]
[1] Queen Mary University of London, UK
[2] University of Cambridge, UK

***Abstract*— Automatic affect analysis and understanding has become a well established research area in the last two decades. However, little attention has been paid to the analysis of the affect expressed in group settings, either in the form of affect expressed by the whole group collectively or affect expressed by each individual member of the group. When it comes to group settings, in addition to affect, analysis of group dynamics is also important for understanding groups, such as group membership, which has an important effect on trustworthiness and cooperation among group members. This PhD project focuses on affect and social dynamic analysis in group settings.**

I. INTRODUCTION

Computational analysis and understanding of groups has been gaining momentum in recent years [1], [2]. The main focus has been on group activity recognition [3], [4] and recognition of the social role of individuals in group settings [5], [6]. However, more recently, other research fields, including emotion recognition, have started to shift their focus from individual to group settings [7], [8].

Emotion communication is an important way through which humans interact socially. Training a system capable of automatically recognizing emotions displayed by humans will advance Human-Computer Interaction further. Therefore, automatic affect analysis has attracted increasing attention and has seen much progress in recent years [9]. However, little attention has been paid to the analysis of the affect expressed by a group of people in a scene or in an interaction setting. From the social psychological perspective, the affect of each individual is influenced by the overall group [10]. From the automatic analysis perspective, Leite *et al.* [11] reported that individual disengagement could be modelled differently in individual and group settings; and their results indicated that more diverse types of disengagement behaviours were shown in group settings than in individual settings. Therefore, on one hand, it would be interesting to study the individual affect expressed in a group setting. However, to the best of our knowledge, most of the existing works on affect analysis in group settings focus on the automatic recognition of collective group-level emotions in static images [12], [8]; and little attention has been paid to the automatic affect analysis of each individual member in a group setting. On the other hand, human behaviours are largely dependent on social context [11], [13]. Specifically, the way humans behave alone is different from the way humans behave in a group setting. Consequently, being alone versus being in-a-group setting may affect the performance and the effectiveness of the automatic analysers that heavily depend on the type of data utilised for training. Pioneering works in this direction have recently emerged in disengagement analysis [11] - the individual disengagement studied in different types of settings (i.e., individual vs. group human-robot interactions). However, little attention has been paid to these diverse settings in the affective computing community.

In addition, research works focusing on the analysis of social dimensions, such as engagement and rapport in group settings have also been introduced [11], [14]. Most of the aforementioned works analyse what is happening within the group. However, there are no works focusing on automatic analysis of the relationship between the members of different groups (e.g., group membership recognition). Group membership recognition refers to recognizing which group each individual belongs to. From a social-psychology perspective, group membership is reported to have an important effect on the trustworthiness and cooperation of group members [15], [16]. Therefore, it would be interesting to investigate automatic recognition of group membership.

II. RESEARCH OBJECTIVES

Automatic analysis and understandings of affect and social dimensions is an active field of research with applications in the field of human-robot interaction, information retrieval, medical diagnosis and so on. The main goal of this research aims to enable the machines to detect and understand people's attitudes, mood and emotions in multi-user environment and to advance human-robot interactions. Most of the previous works focus on the analysis in individual settings with only one single person in an image or a video. However, it is now understood that the degree of variation and effect between individual and group settings is significant (e.g., differences in facial and body behaviours, timing and dynamics) [17], [10], [8]. Therefore, in the past few years, a few works shift to affect analysis in group settings [12], [7], [8]. For automatic analysis in group videos, there are still a multiple problems unaddressed. Previous works, on one hand, all focus on static images rather than dynamic videos, but dynamic videos naturally enable the use of temporal as well as spatial information which are very informative for recognizing human affect and group dynamics. On the other hand, all of the previous works did group-level affect analysis, but did not provide any analysis of the individuals in the group. This PhD project aims at developing algorithms, techniques and applications for affect and dynamics analysis in group videos.

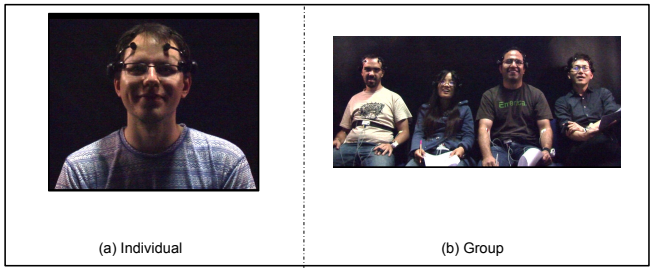

Fig. 1: The setup for individual and group data acquisition.

III. Challenges

To transfer the current affect analysis algorithms to work on data in the group settings, there are several challenges. On one hand, as there are multiple people in the scene or in an interaction, compared to individual settings, there are problems with occlusions, head and body pose variations, illumination variations and interaction taking place between various number of people. The existing methodologies in the field of affective computing mainly focus on extracting generic visual features such as brightness, lighting key and color energy from faces. However, in group settings, due to occlusions, variations in pose and settings, facial information alone are not sufficient for analysing the affect and dynamics. Body and gesture information is of significant importance in affective analysis in group settings. On the other hand, in group setting, the affective states of each individual are not only influenced by the physical environment, but also affected by the other group members and the whole group. Thus, it makes the analysis in group settings more complex. In addition to the affect analysis, in group settings, group dynamics within the whole group or among part of the group members, such as behavioural mimicry and emotional contagion, makes analysis in group settings more challenging than that in individual settings.

IV. Data Collection and Annotation

Two datasets are used in the below works, namely the individual dataset and group dataset. They were collected by another PhD student in our group. Sixteen participants (8 females and 8 males), aged between 25 and 38 were recorded while watching affective movies. Each participant was recorded in both individual and group settings. For the individual dataset, each participant watched sixteen short movies separately. For the group dataset, the participants were arranged into four groups with four participants in each group, watching four long movies together. In order to maximize interactions, groups were formed to include people that knew each other, being friends, colleagues, or people with similar cultural background. Videos were recorded at 1280×720 resolution, 25fps. The setup for these two settings is shown in Fig. 1.

Annotation. Independent observer annotations were obtained from external human labellers who are all researchers working on affect analysis. An in-house emotion annotation tool, that requires the labellers to scroll a bar between a range of values (0 and 1), was used. Individual (short) videos were divided into 10-second clips. All clips were annotated except the first and last 10s of each video. For the group (long) videos, 10-second clips were annotated for every 2 minutes starting from the first minute, e.g., the interval for 00:50∼1:00 min, 2:50∼3:00 min etc. Each labeller was presented with that 10-second clip of each participant separately and was asked to observe the non-verbal behaviours without hearing any audio. A single annotation was given by each labeller after watching one 10-second clip. In order to avoid confusion, arousal and valence annotations were obtained separately.

Analysis of Annotations. As there are three labellers annotating the group videos, the inter-labeller agreement is assessed. Cronbach's α [18] and Fleiss' Kappa [19] statistic, widely used in literature, were computed. The Cronbach's α was calculated directly from the continuous annotations. As Fleiss' Kappa can only be used for the categorical ratings, prior to computing the Fleiss' Kappa, both arousal and valence annotations were first quantised into two classes using the average of all of the annotations as thresholds (i.e., 0.4 for arousal and 0.5 for valence). In this way, arousal is quantised into *high* and *low* and valence is quantised into *positive* and *negative*. After the first annotation round, the Cronbach's α was computed for each subject and the average of all subjects. The displays of subjects with Cronbach's α below the average were reannotated through discussions, and each labeller's annotation was subsequently normalised using Equation 1, where $X = [x_1, x_2, x_3 \ldots x_n]$ refers to all annotations from one labeller, and n is the number of the 10-second recordings.

$$z_i = \frac{x_i - min(X)}{max(X) - min(X)} \tag{1}$$

After the re-annotation and normalization, we obtained 0.95 for arousal and 0.85 for valence with Cronbach's α, and 0.73 for arousal and 0.75 for valence with Fleiss' Kappa, which indicates a very strong inter-labeller reliability.

The average of annotations from all labellers was calculated and used as thresholds (i.e., 0.5 for both arousal and valence in individual dataset, and 0.4 for arousal and 0.5 for valence in group dataset) to quantize the arousal and valence annotations into two classes – *high* and *low* arousal and *positive* and *negative* valence.

V. Proposed Methods

A. Individual Affect Analysis in Group Videos

In [20], a framework was proposed for the prediction of individual emotions and group membership in group videos by multimodal analysis of face and body features. The proposed framework is illustrated in Fig. 2. It aims to investigate the individual affect responses when the participants are watching long-term videos (i.e., 14-24 mins) in group settings. For representing faces, both geometric and appearance features are utilised. Facial landmark trajectories are used as geometric features. In order to encode both spatial and temporal information, a novel volume based Quantised Local Zernike Moment(QLZM) was proposed as facial appearance

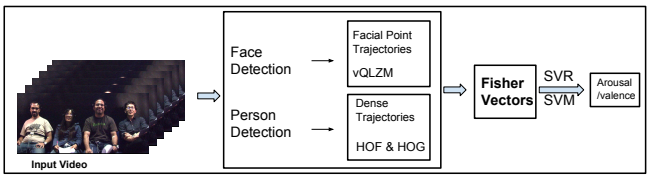

Fig. 2: Illustration of the proposed framework for individual affect analysis.

feature, which is extracted along facial landmark trajectories and illustrated in Figure 3. For representing bodies, dense trajectories are first extracted and then Histogram of Oriented Gradients (HOG) and Histograms of Optical Flow (HOF) descriptors are extracted along the trajectories. Prior to being fed to the classifier and regressor, all of the descriptors are encoded into Fisher Vector (FV) representations. Multiple experiments are carried out to investigate the subject-dependent and subject-independent affect classification and regression using unimodal and multi-modal visual signals. It is found that the proposed volume based QLZM outperformed the other unimodal features.

B. Affect Analysis across Individual and Group Videos

A framework was proposed for the prediction of individual valence and arousal in both individual and group videos by using spatio-temporal facial features in [21]. The framework is illustrated in Fig. 4. Our goal is to investigate whether and how affect recognition differs when training a model using the expressive data of individuals when they are alone versus when they are in a group setting. More specifically, how do the data for the same individual acquired in different settings influence affect recognition? Is it possible to obtain a similar performance on group data using a model trained with individual data and *vice versa*? To this end, three different classification models were trained, including one model trained with data from the individual dataset (i.e., *Individual Model*), one model trained with data from the group dataset (i.e., *Group Model*) and a third model trained with data from both the individual and the group datasets (i.e., *Combined Model*). When training the different models, we use the same feature representation, i.e., facial spatio-temporal representations. Multiple experiments were carried out to investigate the performance of different models on different types of data. The experimental results show that (1) the affect model trained with group data performs better on individual test data than the model trained with individual data tested on group data, indicating that facial behaviours expressed in a group setting capture more variation than in an individual setting; and (2) the combined model does not show better performance than the aect model trained with a specic type of data (i.e., individual or group), but proves a good compromise.

C. Group Membership Recognition

Group membership here refers to recognition which group of each individual is part of. In [22], we propose a novel solution to the group membership recognition problem. We

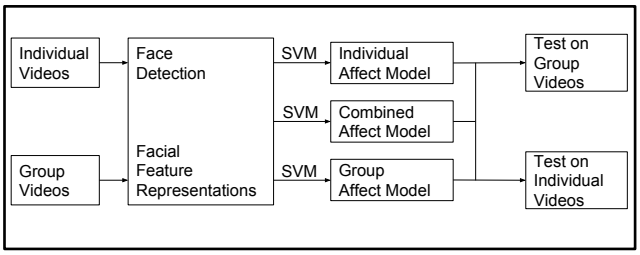

Fig. 4: Illustration of the proposed framework for affect analysis across individual and group videos.

introduce a novel two-phase Support Vector Machine (SVM) based *specific recognition model* that is learned using an optimized *generic recognition model*. More specifically, the data at hand consists of recordings (videos) of different groups watching different movies. Previous work [20] focused on group membership recognition across all different videos, which in our two-phase framework is referred to as the *generic recognition model*. However, we note that group members behave distinctly while watching different movies, which limits the performance of the *generic recognition model*. If we attempt to solve the membership recognition problem with an *independent recognition model* using only samples from the same video, it also becomes very challenging due to the small number of samples available from each video. When group members are watching different movies, they may react differently, however, they are still part of the same setting performing the same task (i.e., sitting in front of the screen watching movies), which enables them to share some common behavioural characteristics. Therefore, we hypothesise that the *generic recognition model* can provide a useful baseline for the optimization of the *specific recognition model* via a two-phase learning. In order to optimize the *specific recognition model*, we first train a *generic recognition model* using all videos and, then, optimize the *specific recognition model* for each specific video based on the optimization results obtained from the *generic recognition model*. The group membership recognition results obtained through this framework show that the proposed *specific recognition model* outperforms both *the generic recognition model* that was trained across all videos using standard linear SVM, and *the independent recognition model* that was trained directly on each video using standard linear SVM.

The framework of the specific group membership recognition model is illustrated in Figure 5 and the independent recognition model is shown in Figure 6.

VI. Research Progress

In the first six months of my PhD, I did a thorough literature review on affect and group dynamics analysis. Then the data was pre-processed and annotated along both arousal and valence. In the second year affect analysis in group videos and affect analysis across individual and group videos were explored. Group membership recognition was

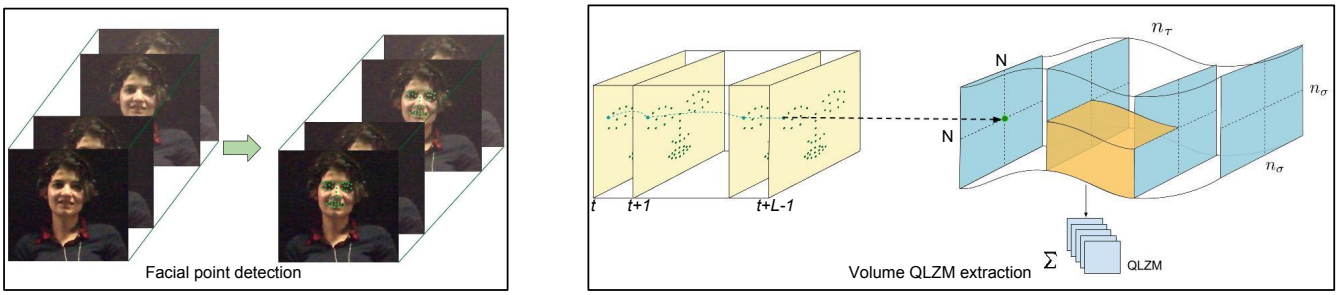

Fig. 3: Illustration of our approach to extract the vQLZM feature.

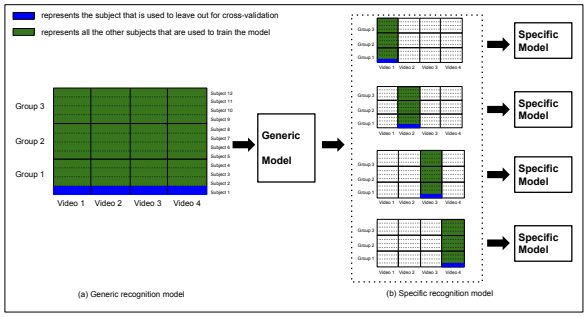

Fig. 5: Illustration of the proposed two-phase learning framework.

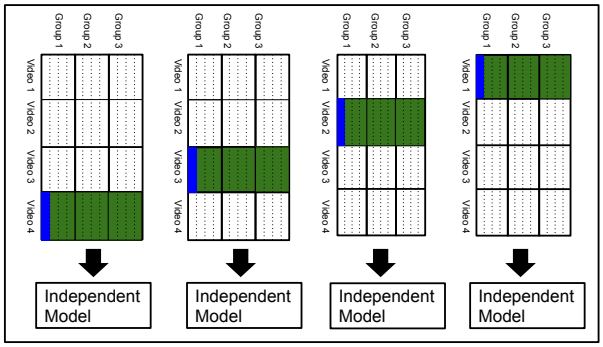

Fig. 6: An illustration of the independent recognition model.

investigated using a two-phase learning model in the third year. Then the two-phase learning model will be investigated further. We will use more data for training / testing and will also apply it to other recognition problems. After that we will have emotional contagion analysis in group videos. Thesis writing will be undertaken in the first half of the fourth year.

VII. ACKNOWLEDGMENTS

The work of Wenxuan Mou is supported by CSC/Queen Mary joint PhD scholarship. The work of Hatice Gunes and Wenxuan Mou is partially funded by the EPSRC under its IDEAS Factory Sandpits call on Digital Personhood (grant ref: EP/L00416X/1).

REFERENCES

[1] M. A. Cronin, L. R. Weingart, and G. Todorova, "Dynamics in groups: Are we there yet?" *The Academy of Management Annals*, 2011.

[2] D. Gatica-Perez, "Automatic nonverbal analysis of social interaction in small groups: A review," *Image and Vision Computing*, 2009.

[3] M. Ibrahim, S. Muralidharan, Z. Deng, A. Vahdat, and G. Mori, "A hierarchical deep temporal model for group activity recognition," *arXiv preprint arXiv:1511.06040*, 2015.

[4] W. Choi, K. Shahid, and S. Savarese, "Learning context for collective activity recognition," in *CVPR*, 2011.

[5] V. Ramanathan, B. Yao, and L. Fei-Fei, "Social role discovery in human events," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[6] A. Gallagher and T. Chen, "Understanding images of groups of people," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[7] A. Dhall, R. Goecke, and T. Gedeon, "Automatic group happiness intensity analysis," *IEEE Trans. on Affective Computing*, 2015.

[8] W. Mou, O. Celiktutan, and H. Gunes, "Group-level arousal and valence recognition in static images: Face, body and context," in *FG*, 2015.

[9] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI)*, 2009.

[10] S. G. Barsade and D. E. Gibson, "Group affect its influence on individual and group outcomes," *Current Directions in Psychological Science*, 2012.

[11] I. Leite, M. McCoy, D. Ullman, N. Salomons, and B. Scassellati, "Comparing models of disengagement in individual and group interactions," in *Proc. ACM/IEEE Int. Conf. Human-Robot Interaction*, 2015.

[12] A. Dhall, J. Joshi, K. Sikka, R. Goecke, and N. Sebe, "The more the merrier: Analysing the affect of a group of people in images," in *FG*, 2015.

[13] R. B. Zajonc *et al.*, *Social facilitation*. Research Center for Group Dynamics, Institute for Social Research, University of Michigan, 1965.

[14] J. L. Hagad, R. Legaspi, M. Numao, and M. Suarez, "Predicting levels of rapport in dyadic interactions through automatic detection of posture and posture congruence," in *ACCV*, 2011.

[15] L. Goette, D. Huffman, and S. Meier, "The impact of group membership on cooperation and norm enforcement: Evidence using random assignment to real social groups," 2006.

[16] M. Williams, "In whom we trust: Group membership as an affective context for trust development," *Academy of management review*, 2001.

[17] S. G. Barsade, "The ripple effect: Emotional contagion and its influence on group behavior," *Administrative Science Quarterly*, 2002.

[18] O. Celiktutan and H. Gunes, "Continuous prediction of perceived traits and social dimensions in space and time," in *ICIP*, 2014.

[19] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, 2008.

[20] W. Mou, H. Gunes, and I. Patras, "Automatic recognition of emotions and membership in group videos," in *CVPRW*, 2016.

[21] ——, "Alone versus in-a-group: A comparative analysis of facial affect recognition," in *Proc. of ACM Int. Conf. on Multimedia*, 2016.

[22] W. Mou, C. Tzelepis, H. Gunes, V. Mezaris, and I. Patras, "Generic to specific recognition models for membership analysis in group videos," in *FG*, 2017.